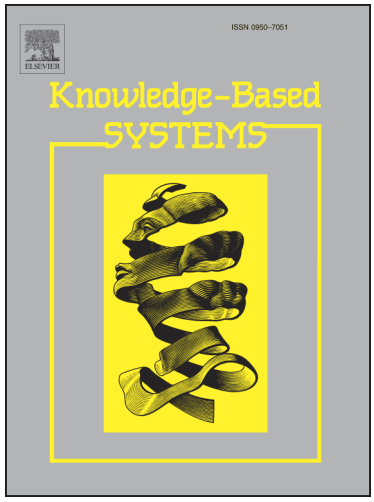

Web Usage Mining with Evolutionary Extraction of Temporal Fuzzy Association Rules

Stephen G. Matthews, Mario A. Gongora, Adrian A. Hopgood, Samad Ahmadi

PII: S0950-7051(13)00281-5
DOI: http://dx.doi.org/10.1016/j.knosys.2013.09.003
Reference: KNOSYS 2629

To appear in: *Knowledge-Based Systems*

Received Date: 16 December 2012
Revised Date: 18 July 2013
Accepted Date: 3 September 2013

Please cite this article as: S.G. Matthews, M.A. Gongora, A.A. Hopgood, S. Ahmadi, Web Usage Mining with Evolutionary Extraction of Temporal Fuzzy Association Rules, *Knowledge-Based Systems* (2013), doi: http://dx.doi.org/10.1016/j.knosys.2013.09.003

Web Usage Mining with Evolutionary Extraction of Temporal Fuzzy Association Rules

Stephen G. Matthews[a], Mario A. Gongora[b], Adrian A. Hopgood[c], Samad Ahmadi[b]

[a]*Intelligent Systems Laboratory, Department of Engineering Mathematics, University of Bristol, Bristol, BS8 1UB, UK*
[b]*Centre for Computational Intelligence, Department of Informatics, De Montfort University, Leicester, LE1 9BH, UK*
[c]*Sheffield Business School, Sheffield Hallam University, Sheffield, S1 1WB, UK*

Abstract

In Web usage mining, fuzzy association rules that have a temporal property can provide useful knowledge about when associations occur. However, there is a problem with traditional temporal fuzzy association rule mining algorithms. Some rules occur at the intersection of fuzzy sets' boundaries where there is less support (lower membership), so the rules are lost. A genetic algorithm (GA)-based solution is described that uses the flexible nature of the 2-tuple linguistic representation to discover rules that occur at the intersection of fuzzy set boundaries. The GA-based approach is enhanced from previous work by including a graph representation and an improved fitness function. A comparison of the GA-based approach with a traditional approach on real-world Web log data discovered rules that were lost with the traditional approach. The GA-based approach is recommended as complementary to existing algorithms, because it discovers extra rules.

1. Introduction

Web usage mining is one type of Web mining (Madria et al., 1999) that attempts to discover patterns of user behaviours that are recorded in the logs of Web servers as users browse Web sites (Cooley et al., 1997b). In this paper, temporal fuzzy association rules are used for Web usage mining. For example, "On a Friday evening, visitors who viewed history.html for a large amount of time also viewed contact-us.html for a medium amount of time". Such rules extend traditional Boolean association rules (Agrawal and Srikant, 1994) by incorporating temporal and fuzzy quantitative features. The temporal feature of the rule is *on a Friday evening*, and the fuzzy features are the *large* and *medium* descriptions.

Matthews et al. (2012) discovered a problem of losing some rules when using traditional methods on synthetic market basket data. Traditional methods follow a two-step process of defining the linguistic labels and membership functions of those labels first, and using them in the mining process. However, the contextual meaning of the linguistic labels can change with events such as seasonal weather, sports games (Saleh and Masseglia, 2010), or unforeseen events, e.g., hurricanes (Leonard, 2005). The problem is that although the meaning can change in a temporal period the membership functions remain the same. For example, a low quantity of ice-cream sales in summer has a different meaning to a low quantity in winter. The membership function does not accurately define the linguistic label for some temporal periods.

Matthews et al. (2012) created a solution that combined the flexibility of the 2-tuple linguistic representation (Herrera and Martínez, 2000) with the search power of a GA. The 2-tuple linguistic representation displaces membership functions laterally along the universe of discourse whilst the linguistic label remains the same. Previous work is improved in this paper by incorporating a graph data structure with an enhanced fitness function. The enhancements enable the approach to work on datasets with real-world complexity in a different domain.

This article is structured as follows: Section 2 provides an overview of related work, Section 3 describes the traditional approach and the original GA-based algorithm, Section 4 introduces enhancements to the GA-based algorithm that was applied to Web log data, Section 5 presents the evaluation of our approach compared with a traditional method, and conclusions are made in Section 6.

2. Related Work

The application of Web usage mining has been categorised as either personalised for learning user profiles, or unpersonalised for user navigation patterns (Srivastava et al., 2000). In this paper, we focus on user navigation patterns represented with fuzzy association rules. Web usage mining can be used for the personalisation of web content, pre-fetching and caching, enhancing Web site design, and customer relationship management in e-commerce (Facca and Lanzi, 2005). Recent work has also applied similar techniques to those in this paper. GAs have mined sequence rules in Web log data (Tuğ et al., 2006) and have also performed subgroup discovery (Carmona et al., 2012). Fuzzy sets have been used to represent the time spent viewing Web pages for fuzzy association rules (Wong et al., 2001) and fuzzy sequence rules (Hong et al., 2002). The temporal and fuzzy features of association rules that are mined in this paper are now reviewed.

The term *temporal* is ambiguous, because it can have different interpretations in temporal data mining (Mitsa, 2010). In this paper, a temporal association rule expresses associations between items from the same transaction, and that association is repeated (occurs frequently) in multiple transactions of a subset of a dataset. For example, a rule may be present in several transactions, and that rule may occur more frequently on a Friday than any other day of that week. Exhibition periods (Lee et al., 2001) are temporal patterns that take into consideration the time when items were introduced into the dataset, e.g., new publications in a publications database. Cyclic patterns (Özden et al., 1998) have rules that occur more frequently in regular periods, such as a rule that occurs every weekend. Temporal patterns with partial periodicity (Han et al., 1998) relax the regularity of cyclic patterns, so the rule may not be present in some cycles of the temporal pattern. These types of temporal association rules are *intra*-transactional, which is different to *inter*-transactional where rules contain items from several transactions spread over a period of time, such as sequence rules (Agrawal and Srikant, 1995).

Quantitative association rule mining extends Boolean association rule mining by discovering rules in quantitative attributes (Srikant and Agrawal, 1996). For example, the time spent viewing a Web page, or the quantities of items sold in a shopping basket. Quantitative association rule mining discretises quantitative attributes into bins. Quantitative association rules suffer from the crisp boundary problem, so fuzzy association rules better deal with unnatural boundaries of crisp intervals (Kuok et al., 1998) and inaccuracies with physical measurements (Chan and Au, 1997). Fuzzy sets (Zadeh, 1965) allow the quantities to be described with linguistic terms (Zadeh, 1975), such as *low* and *high*.

The temporal property of not discovering rare fuzzy itemsets (Weng, 2011) is different to our research, because we focus on how the fuzzy sets are defined instead of only the temporal property. Au and Chan (2002) also mine fuzzy association rules in temporal partitions of the dataset, and they follow the same two-step process, which can lose rules.

3. Temporal Fuzzy Association Rule Mining

Two approaches for mining temporal fuzzy association rules were run on the United States Environmental Protection Agency (EPA) dataset. The purpose is to demonstrate how the flexibility of the 2-tuple linguistic representation approach can help to discover rules on real-world data that a traditional approach cannot. The two approaches are described here, and enhancements to the GA-based approach are explained in Section 4.

3.1. FuzzyApriori

FuzzyApriori (Hong et al., 2001) is an extension to the Apriori algorithm (Agrawal and Srikant, 1994) that uses a breadth-first search. FuzzyApriori uses fuzzy sets to express quantities of items with linguistic terms, but it does not consider any temporal pattern. So, the dataset is partitioned according to its temporal dimension, such as by hour, and FuzzyApriori is executed on each dataset partition separately. The systematic search of the temporal dimension allows for the discovery of temporal features of fuzzy association rules. This is similar to the first approach for mining cyclic association rules (Özden et al., 1998) where the dataset is also partitioned according to the temporal dimension. The rules mined from each dataset partition are aggregated into a final rule set, which is the end result.

Due to the static nature of membership functions in existing approaches, not all temporal fuzzy association rules can be discovered, hence some are lost. Au and Chan (2002) also mine fuzzy association rules in temporal partitions of the dataset, which has been discussed in Section 2. Au and Chan (2002) use a different search method in the two-step process, but in theory the same problem of losing exists, because the fuzzy sets are defined first and they are static. For

this reason, a method based on the seminal Apriori algorithm is only compared, i.e., FuzzyApriori.

3.2. CHC with 2-tuple linguistic representation

The GA-based approach was first described in Matthews et al. (2012), so an overview is given before introducing enhancements in Section 4. The pseudocode of the algorithm is described in Appendix A. The GA-based approach by Matthews et al. is not considered to be traditional like FuzzyApriori, because it is not an exhaustive search method. Instead, a stochastic search method is applied – a GA called Cross-generational elitist selection, Heterogeneous re-combination, and Cataclysmic mutation (CHC) (Eshelman, 1991). The contextual change of meaning for linguistic labels is modelled with the 2-tuple linguistic representation, which is a flexible representation. The crucial difference from other temporal fuzzy association rule mining approaches is that Matthews et al. simultaneously search for membership function parameters and the items in the rule, as well as the temporal period when the rule occurs. This overcomes the problem of membership functions remaining the same when there is a contextual change in the meaning of linguistic labels. Alternative GA-based approaches that simultaneously search for fuzzy sets and rules do exist, but they perform different tasks, i.e., control (Homaifar and McCormick, 1995), classification (Zhou and Khotanzad, 2007), and fuzzy modelling (Delgado et al., 1997). The GA-based approach uses Iterative Rule Learning (IRL) (González and Herrera, 1997). IRL represents one rule in a chromosome. One rule is used from the final population of a GA. More rules are learnt by repeating the GA and penalising previously learnt rules in the fitness function.

4. Enhanced Temporal Fuzzy Association Rule Mining

The GA-based approach is extended with an enhanced fitness function. A weight in the fitness function provides a preference-based multi-objective model to overcome confidence dominating the fitness (Matthews et al., 2012). Previous approaches also use Pareto-based multi-objective models (Matthews et al., 2011), however, selecting a single rule from the Pareto front (for IRL) is a challenging problem. A chromosome C has mixed types, and is defined as $C = (e_l, e_u, i_1, s_1, \alpha_1, a_1, \ldots, i_k, s_k, \alpha_k, a_k)$ where the lower temporal endpoint is e_l (start of time window), the upper temporal endpoint is e_u (end of time window), i is the uniform resource locator (URL), s is the linguistic label expressing the page view time for that URL (e.g., *medium*), α is the lateral displacement of that linguistic label, a determines the antecedent/consequent part, and k is the number of URLs in a rule. For example, a chromosome $(807127200, 807130800, \text{“}/Rules.html\text{”}, \text{“}medium\text{”}, -0.42, antecedent, \ldots, \text{“/”}, \text{“}medium\text{”}, 0.31, consequent)$ represents the rule “IF *view time of /Rules.html* is (*medium*, -0.42) THEN *view time of /* is (*medium*, 0.31) during the period from 807127200 to 807130800” (unixtime). A single rule is represented and extracted from a chromosome, because the lateral displacements of a fuzzy set are specific to each rule.

The fuzzy support count of a chromosome C in a single transaction t_j is defined from Hong et al. (2001) as

$$\text{FuzSupTran}(C^{(t_j)}) = \min_{n=1}^{k} \mu_{(s_n,\alpha_n)}(t_j^{(i_n)}), \tag{1}$$

where μ is the degree of membership for a linguistic label s_n and lateral displacement α_n for item i_n with a rule of length k and for one transaction t_j where j is a dataset transaction ID (TID). The minimum is used for intersection of all the clauses, which is the same method of intersection used in FuzzyApriori.

FuzSupTran is then used to calculate fuzzy support counts across multiple transactions and the fitness is defined as

$$\text{Fitness}(C) = \left(\frac{\sum_{j=e_l}^{e_u} \text{FuzSupTran}(C_X^{(t_j)} \cap C_Y^{(t_j)})}{e_u - e_l} \right) + w \left(\frac{\sum_{j=e_l}^{e_u} \text{FuzSupTran}(C_X^{(t_j)} \cap C_Y^{(t_j)})}{\sum_{j=e_l}^{e_u} \text{FuzSupTran}(C_X^{(t_j)})} \right), \tag{2}$$

where C is a chromosome, X is the rule antecedent, Y is the rule consequent, j is a dataset TID from the e_l lower endpoint to the e_u upper endpoint, and w is a weight applied to the confidence measure. Hence, $C_X^{(t_j)}$ and $C_Y^{(t_j)}$ are the rule antecedent and the rule consequent respectively for one transaction in the dataset. A weight is required to avoid local minima that occur as a result of the magnitude of confidence being higher than the magnitude of temporal fuzzy support; a GA produces high confidence values (Matthews et al., 2012; Alcala-Fdez et al., 2010) compared with a smaller magnitude of support values. For example, a temporal fuzzy support value of 0.001 is smaller than a confidence value of 0.1, so the confidence value has more influence than the temporal fuzzy support. The weight was determined

from multiple runs of the algorithm so that the temporal fuzzy support and weighted confidence had the same order of magnitude.

We also extend our previous work with a graph representation to enhance the efficiency of searching for URLs/items in a rule. Our previous approach allowed the generation of invalid chromosomes with rules that did not exist in the dataset. Such rules are detrimental to the search process, and are undesirable in the final rule set.

The dataset is transformed from rows and columns to a cyclical undirected graph. The purpose is to ensure that chromosomes contain valid itemsets that are present in the dataset and also to reduce the itemset search space. The tabular representation is used for fitness evaluation, and the graph representation is used during initialisation and crossover.

An undirected graph G is a pair of finite sets (V, E) where V is a non-empty set of vertices and E is a set of pairs (e, t). Each pair in E consists of an edge e and a non-empty finite set of TIDs t. Each edge e is an unordered pair of vertices (a, b). The definition extends regular graphs by including a set of TIDs for each edge.

An example is presented to demonstrate the construction of the graph. Table 1 is a small example of a quantitative dataset transformed into the graph of Figure 1. Each edge represents the co-occurrence of two items. Items are vertices. The TIDs of the co-occurrence are also on an edge. Edges are paired with a set of TIDs to identify the co-occurrence of items. If there is no set of TIDs for an edge then an edge does not exist.

Table 1: Example dataset containing three items/URLs (A, B and C) with quantities for four transactions in vertical layout

TID	A	B	C
1	4	6	12
2	0	2	14
3	16	11	0
4	1	0	13

The vertices for the example graph are $V = \{A, B, C\}$, and the edges are $E = ((A, B), \{1, 3\})$, $((B, C), \{1, 2\})$, $((A, C), \{1, 4\})$, $((A, A), \{1, 3, 4\})$, $((B, B), \{1, 2, 3\})$, $((C, C), \{1, 2, 4\})\}$. A loop connects a vertex to itself. These edges are shown in Figure 1 as lines that loop to the same vertex, i.e., TIDs $\{1, 3, 4\}$ for vertex A, TIDs $\{1, 2, 3\}$ for vertex B, and TIDs $\{1, 2, 4\}$ for vertex C.

The graph representation is incorporated into initialisation and crossover of chromosomes. The algorithms

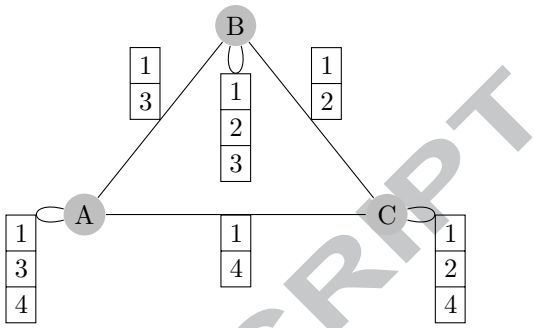

Figure 1: Example graph transformed from dataset in Table 1

are defined in Appendix A. The algorithm Hybrid-Crossover prevents crossover from producing invalid itemsets. Potential offspring are first checked to identify if the resulting itemsets are present in the specific temporal partitions of the dataset. If the resulting offspring are not present then the items are not swapped. Algorithm CheckGraph in Appendix A uses the graph data structure to ensure offspring are valid itemsets in a temporal period.

5. Evaluation

The dataset and methodology for analysing the enhanced GA-based approach are discussed and results are then presented.

5.1. Data

A Web log dataset has both temporal and quantitative features. The temporal feature is the timestamp of a request made to the server, and the quantitative feature is the page view time in seconds.

The EPA dataset[1] is a collection of Hypertext Transfer Protocol (HTTP) requests to a Web server collected from a 24-hour period. The geographical location of the Web server is Research Triangle Park, NC, USA. The EPA dataset was recorded from 23:53:25 29th August 1995 EDT to 23:53:07 30th August 1995 EDT. The EPA dataset has 47748 requests: 46014 GET requests, 1622 POST requests, 107 HEAD requests, and 6 invalid requests. Table 2 shows a sample of records from the EPA dataset before cleaning and preprocessing.

The EPA dataset was cleaned by removing all records assumed to be the Web site's design or a non-traversable Web page (suffixes: gif, xbm, zip, pdf, exe, gz, wpd,

[1] Available from The Internet Traffic Archive (`http://ita.ee.lbl.gov/`)

Table 2: The first 4 records from the EPA dataset

Host	Date	Request	HTTP reply code	Bytes in reply
141.243.1.172	[29/Aug/1995:23:53:25]	"GET /Software.html HTTP/1.0"	200	1497
query2.lycos.cs.cmu.edu	[29/Aug/1995:23:53:36]	"GET /Consumer.html HTTP/1.0"	200	1325
tanuki.twics.com	[29/Aug/1995:23:53:53]	"GET /News.html HTTP/1.0"	200	1014
wpbfl2-45.gate.net	[29/Aug/1995:23:54:15]	"GET / HTTP/1.0"	200	4889

wp, dct, jpg, and imf). All records that did not have a GET request method were removed. After preprocessing, there were 2688 transactions and 5147 URLs. Preprocessing consisted of:

1. A 10-minute time window was used Cooley et al. (1997a), which assumes that visitors do not view the page for more than 10 minutes.
2. Maximal forward reference transaction identification (Chen et al., 1996) produced lists of URLs, which are referred to as transactions.
3. Some resulting transactions contained the same URLs next to each other. This is likely to be caused by refreshing the Web page, so subsequent occurrences of the same URL were removed.
4. Two URL requests are required to determine the page view time of a URL by calculating the difference in timestamps. For example, history.html accessed at 12:00:10 and contact-us.html accessed at 12:00:30 has a view time of 20 seconds. Transactions with 2 or fewer URLs were removed to ensure that a page view time can be calculated.
5. For FuzzyApriori, the dataset was partitioned by hour.

5.2. Methodology

The methodology for evaluating the approaches on the real-world EPA dataset is described. The aim of the evaluation is to identify whether the GA-based approach can discover rules that the traditional approach cannot. This is a novel approach to traditional methodologies because the focus is on discovering lost rules on a real-world dataset, which is a new unrecognised problem that warrants a different methodology (Matthews et al., 2012). The methodology for evaluation is the same as Matthews et al. (2012), but the analysis of the results is simplified to improve clarity.

The linguistic labels and membership functions are defined first. The traditional and GA-based approaches were run using the same linguistic labels and membership functions. The result was two sets of temporal fuzzy association rules: one set containing the traditional fuzzy set representation, and the other containing the 2-tuple linguistic representation. The two sets of rules were compared to identify rules that matched and rules that did not match.

The method of rule comparison from IRL was used (Matthews et al., 2012). Each clause of the rule is compared. If the items/URLs and linguistic labels of two clauses match, then the lateral displacements are compared. The lateral displacements are considered to be the same if the difference in absolute values of lateral displacements is less than a lateral displacement threshold of 0.5. For example, for a lateral displacement threshold of 0.5 and two lateral displacements, -0.45 and -0.05, the absolute difference is 0.4 so the fuzzy sets are considered to be the same.

5.3. Results

The two approaches for discovering temporal fuzzy association rules were run and the rules were compared. Results of the comparison and an example of a lost rule are presented here. The algorithms were implemented in Java within the KEEL tool (Knowledge Extraction based on Evolutionary Learning) (Alcalá-Fdez et al., 2009). The experiments were conducted on a personal computer with a 64-bit 2GHz dual-core processor and 3GB RAM. FuzzyApriori had a minimum temporal fuzzy support of 0.0011 and a minimum confidence of 0.5. Rules are discarded because their measures fall below either the minimum temporal fuzzy support or the minimum confidence. The minimum confidence was not set high, so that rules are not discarded because of *low* confidence when the rules have *high* temporal fuzzy support. The reason for the minimum confidence value is that the temporal fuzzy support is a key factor in a temporal pattern. Furthermore, the minimum confidence was increased from 0.05 in our previous approach (Matthews et al., 2012). The population size was 50, and the PCBLX crossover operator parameter was 1 (Matthews et al., 2012). IRL was configured to produce the same percentage of rule lengths as

FuzzyApriori. For example, if FuzzyApriori produced 50% with length 2 and 50% with length 3, then IRL produced 50% with length 2 and 50% with length 3.

Initially, the GA-based approach was run once to assess a typical run. Table 3 shows statistics of both approaches. The GA-based approach was limited to 100 rules and the systematic search with FuzzyApriori discovered 762 rules. The arithmetic mean of temporal fuzzy support was higher for the GA-based approach, but it had a lower confidence value. The differences in distributions of both measures are shown in Figures 2 and 3. The GA-based approach takes longer, but this is outweighed by its benefit of discovering rules with higher temporal fuzzy support.

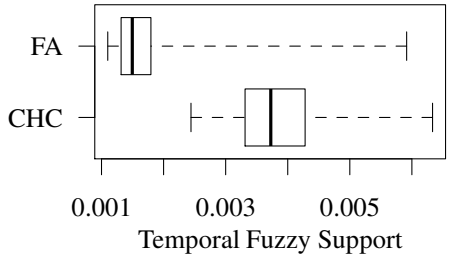

Figure 2: Boxplot of temporal fuzzy support for FuzzyApriori (FA) and one run of CHC.

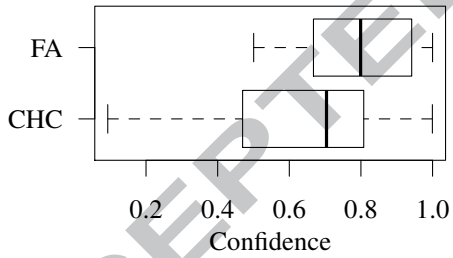

Figure 3: Boxplot of confidence for FuzzyApriori (FA) and one run of CHC.

The evolution of the best fitness for one run of CHC (one iteration of IRL) is shown in Figure 4. It can be observed that the best fitness increases through the generations. The large jumps in best fitness are likely to be caused by a change in nominal data (e.g., item, linguistic label) in the chromosome rather than interval data (e.g., lateral displacement).

The method of evaluation from Matthews et al. (2012), which is stated in Section 5, is now used. The purpose is to identify what rules the GA-based approach can discover that a traditional approach cannot discover in Web log data. The experiments that follow were conducted by running the GA-based approach 30 times and the percentages are arithmetic means of

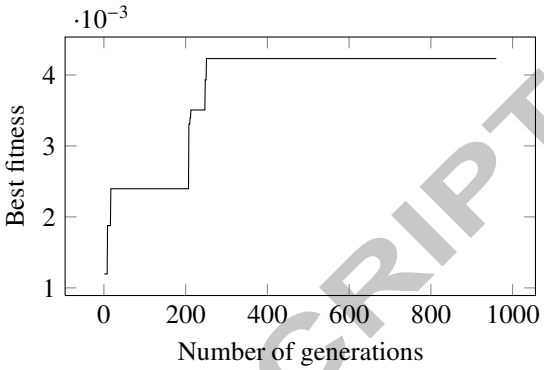

Figure 4: Best fitness during one run of CHC.

all runs. Tables 4 and 5 show that 49.23% of the 100 rules discovered with the GA-based approach were also discovered with FuzzyApriori. Of those 49.23% from the GA-based approach, 2.93% had a reduction in temporal fuzzy support and 46.30% had an increase in temporal fuzzy support. When analysing confidence of the same 49.23% of rules, 9.20% had a reduction, 0.30% did not change and 39.73% had an increase. The GA-based approach has rediscovered 49.23% of rules that were already discovered by FuzzyApriori. However, the quality increased significantly for the majority of these rules. Over half the rules (50.77%) were only discovered with the GA-based approach. These 50.77% were not discovered by FuzzyApriori because they fell below one or both thresholds.

Table 4: Analysis of temporal fuzzy support for rules discovered in CHC and FuzzyApriori (FA). Percentages show how the GA-based approach changed the measure with a decrease (-ve(%)), no change (0(%)), and an increase (+ve(%)).

	Arithmetic mean of change in Temporal Fuzzy Support			
	-ve(%)	0(%)	+ve(%)	Total(%)
CHC and FA	2.93	0.00	46.30	49.23
CHC only	0.00	0.00	50.77	50.77

It is important to understand which of the 50.77% rules in the EPA dataset are now above the thresholds. When a rule is above both thresholds, it is deemed to be significant. Table 6 shows the percentage of rules that were below the threshold(s), and the percentage of rules now above the threshold(s). Rules now above the thresholds are significant to this research because the GA-based approach has learnt the lateral displacement of membership functions so that a rule is now above the threshold(s).

Table 3: Results for FuzzyApriori and one run of CHC

Measure	CHC	FuzzyApriori
Number of Rules	100	782
Arithmetic mean of temporal fuzzy support (4 s.f.)	0.0039	0.0017
Arithmetic mean of confidence (4 s.f.)	0.6078	0.7918
Execution time (minutes)	1422.65	1.52

Table 6: Rules below threshold and rules above threshold

	Discarded by threshold(s) (%)	Greater than or equal to threshold(s) (%)
Below min. temporal fuzzy support only	8.30	8.30
Below min. confidence only	27.03	5.90
Below both (above both)	15.43	11.13
Total	50.77	25.33

Table 5: Analysis of confidence for rules discovered in CHC and FuzzyApriori (FA). Percentages show how the GA-based approach changed the measure with a decrease (-ve(%)), no change (0(%)), and an increase (+ve(%)).

	Arithmetic mean of change in Confidence			
	-ve(%)	0(%)	+ve(%)	Total(%)
CHC and FA	9.20	0.30	39.73	49.23
CHC only	9.33	0.24	41.20	50.77

The results are reported using percentage, which is a relative measure of the 100 rules. It is important to note that increasing the number of rules in IRL may not discover more lost rules than those discovered in the 100 rules. In such case, the percentage would decrease. However, lost rules are still discovered, and it may only be one rule that is of great significance/interest to the user.

An example of a temporal fuzzy association rule is presented below. The rule *was not* discovered with FuzzyApriori, because the temporal fuzzy support of 0.0005 was below the threshold of 0.0011, and the confidence of 0.44 was below the threshold of 0.5.

Endpoints (unixtime): 807127200–807130800
Rule: IF *view time of /Rules.html* is *medium*
THEN *view time of /* is *medium*
Temporal Fuzzy Support: 0.0005
Confidence: 0.44

The same example rule *was* discovered with the GA-based approach, as shown below, but with lateral displacements from the 2-tuple linguistic representation. The rule demonstrates knowledge that was lost with a traditional approach, but learnt with the GA-based approach.

Endpoints (unixtime): 807127200–807130800
Rule: IF *view time of /Rules.html* is (*medium*, -0.49)
THEN *view time of /* is (*medium*, -0.49)
Temporal Fuzzy Support: 0.004
Confidence: 0.67

Further experiments were conducted with CHC and FuzzyApriori on a synthetic market basket dataset to examine scalability and parameter settings. Preliminary results showed that increasing the number of transactions from 10,000 to 90,000 transactions (same as Web site visitors) produced a constant number of lost rules and execution time is linear. And, increasing the number of items from 1000 to 5000 items (same as URLs) decreased the number of lost rules and the execution time is linear. Full experimentation and statistical analysis on more real-world examples are subject of our ongoing future work.

6. Conclusions

We have demonstrated the problem of losing temporal fuzzy association rules on real-world Web log data for the first time and presented a novel solution. Our previous approach of using a GA and the 2-tuple linguistic representation has been improved by transforming the dataset to a graph, which ensures valid itemsets are discovered, and modifying the fitness function.

The execution time of the GA-based approach is longer, however, the contribution to knowledge is that it can discover rules that a traditional approach cannot, and the rules have higher temporal fuzzy support. The GA-based approach is recommended as complementary to existing algorithms, because it discovers extra rules that a traditional algorithm does not. The decision to use this complementary approach can rely on understanding what temporal changes may be present in the application domain (e.g., seasonal and/or scheduled events).

It is important to note that lowering minimum support/confidence would overcome the problem of losing rules with traditional approaches, however, the number of rules increases, which is undesirable in association rule mining. Further work will explore different enhancements, and different approaches to tackle the same problem.

Appendix A. Algorithms

Algorithm 1. IRL with CHC

Begin

While maximum number of iterations not reached **do**

Generate initial population
Evaluate initial population and initialise L
While maximum number of fitness evaluations not reached **do**

Select individuals from parents
Recombine individuals to form offspring
Evaluate offspring
Combine offspring with parents, and select the best N individuals for the next population.
If there are no new individuals, or the best chromosome does not change, then $L = L-1$.
If $L < 0$ then reinitialise the population

End (While)
Add best rule to final rule set

End (While)

End

Algorithm 2. ***HybridCrossover***

Inputs:

$k \leftarrow$ Length of rule;
$P \leftarrow$ Two parent chromosomes;

Outputs:

O; // Two offspring chromosomes

Begin

$n \leftarrow 0$; // Initialise loop variable to first index
$O \leftarrow P$; // Create offspring from identical copies of parents;
If offspring have matching items **Then**

Move clauses, containing matching items, to same loci;

End (If)
While $n < k$ **do** // Loop through every index in rule

If O items are identical **AND** linguistic labels are identical **Then**

Uniform crossover of lateral displacement using parent centric BLX-α (PCBLX-α);
Uniform crossover of antecedent-consequent parameter using swap;

End (If)
If O items are identical **AND** linguistic labels are not identical **Then**

Uniform crossover of {linguistic label, lateral displacement} using swap;
Uniform crossover of antecedent-consequent parameter using swap;

End (If)
If O items are not identical **Then**

If O_1 is present in endpoints of O_2 using Algorithm CheckGraph **AND** O_2 is present in endpoints of O_1 using Algorithm CheckGraph **Then**

Uniform crossover of {item, linguistic label, lateral displacement} using swap;
Uniform crossover of antecedent-consequent parameter using swap;

End (If)

End (If)
$n \leftarrow n + 1$; // Increment loop variable

End (While)

End

Algorithm 3. ***CheckGraph***

Inputs:

$k \leftarrow$ Length of itemset to be checked;

$I \leftarrow$ Itemset to be checked;

$j \leftarrow$ Candidate item from itemset I that is to be checked;

$M \leftarrow$ Adjacency matrix of graph of dataset;

$(e_l, e_u) \leftarrow$ Lower and upper endpoints of temporal period;

Outputs:

TRUE or FALSE

Begin

$n \leftarrow 0$; // Initialise loop variable to first index

$T \leftarrow \emptyset$; // Initialise set of transactions to the empty set

$p \leftarrow$ Clause index of candidate item j;

While $n < k$ **do** // Loop through every index in itemset

// If index of candidate item not equal to current index

If $n \neq p$ **Then**

// If T used for first time

If $n = 0$ **OR** $p = 0$ **Then**

// Initialise

$T \leftarrow M_{I_n,j}$ that are >= e_l and < e_u;

Else

$T \leftarrow T \cap M_{I_n,j}$; // Update transaction IDs set with transactions containing candidate item (j) and current item (I_n)

End (If)

End (If)

$n \leftarrow n + 1$; // Increment loop variable

End (While)

If $T = \emptyset$ **Then** return FALSE;

Else return TRUE;

End (If)

End

References

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile, Morgan Kaufmann Publishers Inc., San Fransisco. pp. 487–499.

Agrawal, R., Srikant, R., 1995. Mining sequential patterns, in: Proceedings of the Eleventh International Conference on Data Engineering, Taipei, Taiwan. pp. 3–14.

Alcala-Fdez, J., Flugy-Pape, N., Bonarini, A., Herrera, F., 2010. Analysis of the effectiveness of the genetic algorithms based on extraction of association rules. Fundamenta Informaticae 98, 1–14.

Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M., Ventura, S., Garrell, J., Otero, J., Romero, C., Bacardit, J., Rivas, V., Fernández, J., Herrera, F., 2009. KEEL: A software tool to assess evolutionary algorithms to data mining problems. Soft Computing - A Fusion of Foundations, Methodologies and Applications 13, 307–318.

Au, W.H., Chan, K., 2002. Fuzzy data mining for discovering changes in association rules over time, in: Proceedings of the 2002 IEEE International Conference on Fuzzy Systems, Honolulu, HI, USA, IEEE, Piscataway. pp. 890–895.

Carmona, C.J., Ramírez-Gallego, S., Torres, F., Bernal, E., del Jesus, M.J., García, S., 2012. Web usage mining to improve the design of an e-commerce website: Orolivesur.com. Expert Systems with Applications 39, 11243–11249.

Chan, K.C.C., Au, W.H., 1997. Mining fuzzy association rules, in: Proceedings of the Sixth International Conference on Information and Knowledge Management, Las Vegas, NV, USA, ACM, New York. pp. 209–215.

Chen, M.S., Park, J.S., Yu, P., 1996. Data mining for path traversal patterns in a web environment, in: Proceedings of the 16th International Conference on Distributed Computing Systems, Hong Kong, IEEE Computer Society, Washington. pp. 385–392.

Cooley, R., Mobasher, B., Srivastava, J., 1997a. Grouping web page references into transactions for mining world wide web browsing patterns, in: Proceedings of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop, Newport Beach, CA, USA, IEEE Computer Society, Washington. pp. 2–9.

Cooley, R., Mobasher, B., Srivastava, J., 1997b. Web mining: information and pattern discovery on the world wide web, in: Proceedings of the Ninth IEEE International Conference on Tools with Artificial Intelligence, Newport Beach, CA, USA, IEEE, Piscataway. pp. 558–567.

Delgado, M., Gómez-Skarmeta, A., Martín, F., 1997. A fuzzy clustering-based rapid prototyping for fuzzy rule-based modeling. IEEE Transactions on Fuzzy Systems 5, 223–233.

Eshelman, L.J., 1991. The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination, in: Foundations of Genetic Algorithms, Morgan Kaufmann. pp. 265–283.

Facca, F.M., Lanzi, P.L., 2005. Mining interesting knowledge from weblogs: a survey. Data & Knowledge Engineering 53, 225–241.

González, A., Herrera, F., 1997. Multi-stage genetic fuzzy systems based on the iterative rule learning approach. Mathware & Soft Computing 4, 233–249.

Han, J., Gong, W., Yin, Y., 1998. Mining segment-wise periodic patterns in time-related databases, in: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, AAAI, Menlo Park. pp. 214–218.

Herrera, F., Martínez, L., 2000. A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Transactions on Fuzzy Systems 8, 746–752.

Homaifar, A., McCormick, E., 1995. Simultaneous design of

membership functions and rule sets for fuzzy controllers using genetic algorithms. IEEE Transactions on Fuzzy Systems 3, 129–139.

Hong, T.P., Kuo, C.S., Chi, S.C., 2001. Trade-off between computation time and number of rules for fuzzy mining from quantitative data. International Journal of Uncertainty, Fuzziness & Knowledge-Based Systems 9, 587–604.

Hong, T.P., Lin, K.Y., Wang, S.L., 2002. Mining linguistic browsing patterns in the world wide web. Soft Computing - A Fusion of Foundations, Methodologies and Applications 6, 329–336.

Kuok, C.M., Fu, A., Wong, M.H., 1998. Mining fuzzy association rules in databases. SIGMOD Record 27, 41–46.

Lee, C.H., Lin, C.R., Chen, M.S., 2001. On mining general temporal association rules in a publication database, in: Proceedings IEEE International Conference on Data Mining, San Jose, CA, USA, IEEE Computer Society, Washington. pp. 337–344.

Leonard, D., 2005. After katrina: Crisis management, the only lifeline was the wal-mart. FORTUNE Magazine (October 3) .

Madria, S.K., Bhowmick, S.S., Ng, W.K., Lim, E.P., 1999. Research issues in web data mining, in: Proceedings of the First International Conference on Data Warehousing and Knowledge Discovery, Florence, Italy, Springer-Verlag, London. pp. 303–312.

Matthews, S.G., Gongora, M.A., Hopgood, A.A., 2011. Evolving temporal fuzzy association rules from quantitative data with a multi-objective evolutionary algorithm, in: Corchado, E., Kurzynski, M., Wozniak, M. (Eds.), Proceedings of the 6th International Conference on Hybrid Artificial Intelligence Systems (HAIS 2011). Springer Berlin / Heidelberg. volume 6678 of *Lecture Notes in Computer Science*, pp. 198–205.

Matthews, S.G., Gongora, M.A., Hopgood, A.A., Ahmadi, S., 2012. Temporal fuzzy association rule mining with 2-tuple linguistic representation, in: Proceedings of the 2012 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2012), Brisbane, Australia, IEEE, Piscataway. pp. 1–8.

Mitsa, T., 2010. Temporal Data Mining. CRC Press Online.

Özden, B., Ramaswamy, S., Silberschatz, A., 1998. Cyclic association rules, in: Proceedings of the Fourteenth International Conference on Data Engineering, Orlando, FL, USA, IEEE Computer Society, Washington. pp. 412–421.

Saleh, B., Masseglia, F., 2010. Discovering frequent behaviors: time is an essential element of the context. Knowledge and Information Systems 28, 1–21.

Srikant, R., Agrawal, R., 1996. Mining quantitative association rules in large relational tables, in: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, ACM, New York. pp. 1–12.

Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N., 2000. Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explorations Newsletter 1, 12–23.

Tuğ, E., Şakiroğlu, M., Arslan, A., 2006. Automatic discovery of the sequential accesses from web log data files via a genetic algorithm. Knowledge-Based Systems 19, 180 – 186.

Weng, C.H., 2011. Mining fuzzy specific rare itemsets for education data. Knowledge-Based Systems 24, 697–708.

Wong, C.K.P., Shiu, S.C.K., Pal, S.K., 2001. Mining fuzzy association rules for web access case adaptation, in: Workshop Proceedings of Soft Computing in Case-Based Reasoning Workshop, in conjunction with the 4th International Conference in Case-Based Reasoning, Vancouver, Canada, Springer-Verlag, London. pp. 213–220.

Zadeh, L.A., 1965. Fuzzy sets. Information Control 8, 338–353.

Zadeh, L.A., 1975. The concept of a linguistic variable and its application to approximate reasoning. Parts I, II, III. Information Sciences 8–9, 199–249, 301–357, 43–80.

Zhou, E., Khotanzad, A., 2007. Fuzzy classifier design using genetic algorithms. Pattern Recognition 40, 3401–3414.